AMENDMENTS TO THE CLAIMS

1. (Currently Amended) A program storage device readable by a machine, tangibly embodying a program of instructions executable by the machine to perform a method ~~steps~~ for speech synthesis that allows user specified pronunciations, the method ~~steps~~ comprising:

~~providing a text string comprising a plurality of words and phonemes and a corresponding spoken audio signal wherein a user specifies a pronunciation of the text string;~~

~~extracting acoustic feature data from said audio signal,~~

aligning ~~the~~ a text string comprising a plurality of words and phenomes and ~~the~~ a user specified spoken audio signal corresponding to a desired pronunciation of the text string ~~acoustic feature data and outputting a set of duration contours indicative of the duration of each word and phoneme~~;

extracting ~~pitch contour~~ prosodic parameters from said spoken audio ~~spoken input~~ signal;

automatically generating a marked-up text corresponding to the spoken ~~utterance~~ audio signal using the prosodic parameters ~~pitch and duration contours~~; and

generating a synthetic waveform using the marked-up text.

2-3. (Cancelled)

4. (Previously Presented) The program storage device of claim 1, wherein the instructions for aligning comprise instructions for segmenting said spoken audio signal into time-segmented regions, wherein each time-segmented region is mapped to a corresponding phoneme.

5. (Previously Presented) The program storage device of claim 1, wherein the alignment is performed using a Viterbi alignment process.

6. (Canceled)

7. (Currently Amended) The program storage device of claim 1, wherein the instructions for automatically generating a marked-up text comprise instruction for directly specifying at least one portion of ~~the pitch~~ duration contours ~~and duration~~ or prosodic parameters as attribute values for mark-up elements.

8. (Currently Amended) The program storage device of claim 1, wherein the instructions for automatically generating a marked-up text comprise instructions for assigning abstract labels to at least one portion of ~~the pitch~~ duration contours ~~and duration~~ or prosodic parameters to generate a high-level markup.

9. (Original) The program storage device of claim 1, wherein the marked-up text is generated using SSML (speech synthesis markup language).

10. (Currently Amended) The program storage device of claim 1, further comprising instruction for processing phonetic content of the spoken ~~utterance~~ audio signal to generate the synthetic waveform having a desired pronunciation.

11. (Currently Amended) A method for speech synthesis that allows user specified pronunciations, the method comprising ~~the steps of~~:

~~providing a text string comprising a plurality of words and phonemes and a corresponding spoken audio signal wherein a user specifies a pronunciation of the text string;~~

~~extracting acoustic feature data from said audio signal;~~

aligning ~~the~~ a text string comprising a plurality of words and phenomes and ~~the~~ a user specified spoken audio signal corresponding to a desired pronunciation of the text string ~~acoustic feature data and outputting a set of duration contours indicative of the duration of each word and phoneme~~;

extracting ~~pitch contour~~ prosodic parameters from said spoken audio ~~spoken input~~ signal;

automatically generating a marked-up text corresponding to the spoken ~~utterance~~ audio signal using the prosodic ~~pitch contour and duration~~ parameters; and

generating a synthetic waveform using the marked-up text.

12-13. (Canceled)

14. (Currently Amended) The method of claim 11, wherein aligning comprises extracting acoustic feature data from the spoken ~~utterance~~ audio signal and time–aligning the spoken ~~input~~ audio signal to the ~~corresponding~~ text string using the acoustic feature data.

15. (Previously Presented) The method of claim 11, wherein aligning is performed using a Viterbi alignment process.

16. (Canceled)

17. (Currently Amended) The method of claim 11, wherein automatically generating a marked-up text comprises directly specifying at least one portion of ~~the pitch~~ duration contours ~~and duration~~ or prosodic parameters as attribute values for mark-up elements.

18. (Currently Amended) The method of claim 11, wherein automatically generating a marked-up text comprises assigning abstract labels to at least one portion of ~~the pitch~~ duration contours ~~and duration~~ or prosodic parameters to generate a high-level markup.

19. (Original) The method of claim 11, wherein the marked-up text is generated using SSML (speech synthesis markup language).

20. (Currently Amended) The method of claim 11, further comprising processing phonetic content of the spoken ~~utterance~~ audio signal to generate the synthetic waveform having a desired pronunciation.

21. (Currently Amended) A text-to-speech (TTS) system that allows user specified pronunciations, comprising:

a prosody analyzer for determining prosodic parameters of a spoken ~~utterance~~ audio signal corresponding to a desired pronunciation of an input text string and automatically generating a marked-up text corresponding to the spoken audio signal ~~utterance~~ using the prosodic parameters ~~wherein a user specifies a pronunciation of the text string with said spoken utterance~~, wherein the prosody analyzer comprises:

~~an acoustic feature extraction module that extracts acoustic feature data from said spoken utterance;~~

an alignment module for aligning the input text string with the spoken ~~utterance~~ audio signal ~~using said acoustic feature data to generate duration contour information of elements comprising the input text string~~[[;]],

a ~~pitch contour~~ prosodic parameter extraction module for determining ~~pitch contour~~ prosodic parameter information for the spoken ~~utterance~~ audio signal [[;]], and

a conversion module for including markup in the input text string ~~in accordance with~~ using the ~~duration and pitch contour information~~ prosodic parameter information to generate ~~the~~ marked-up text; and

a TTS system for generating a synthetic waveform using the marked-up text.

22. (Currently Amended) The system of claim 21, further comprising a user interface that enables a user to input the spoken ~~utterance~~ audio signal and input a text string corresponding to the spoken ~~utterance~~ audio signal.

23. (Currently Amended) The system of claim 21, wherein the prosody analyzer processes phonetic content of the spoken ~~utterance~~ audio signal to generate the synthetic waveform having a desired pronunciation.

24-28. (Canceled)

29. (Currently Amended) The program storage device of claim ~~1~~ 33, wherein extracting acoustic feature data from said spoken audio signal comprises digitizing the spoken audio signal into a set of frames and transforming ~~the~~ digitized input waveforms into a set of feature vectors on a frame-by-frame basis.

30. (Previously Presented) The program storage device of claim 29, wherein transforming the digitized input includes producing a 24-dimensional cepstra feature vector for every 10ms of the spoken audio signal, concatenating frames to the left and to the right of a current frame to augment a current cepstral vector, and reducing each augmented cepstral vector to a 60-dimensional feature vector using linear discriminant analysis.

31. (Currently Amended) The method of claim ~~11~~ 34, wherein extracting acoustic feature data from said spoken audio signal comprises digitizing the spoken audio signal into a set of frames and transforming ~~the~~ digitized input waveforms into a set of feature vectors on a frame-by-frame basis.

32. (Previously Presented) The method of claim 31, wherein transforming the digitized input includes producing a 24-dimensional cepstra feature vector for every 10ms of the spoken audio signal, concatenating frames to the left and to the right of a current frame to augment a current cepstral vector, and reducing each augmented cepstral vector to a 60-dimensional feature vector using linear discriminant analysis.

33. (New) The program storage device of claim 1 wherein the method further comprises extracting acoustic feature data from said spoken audio signal and wherein the aligning further comprises outputting a set of duration contours.

34. (New) The method of claim 11 further comprising extracting acoustic feature data from said spoken audio signal and wherein the aligning further comprises outputting a set of duration contours.

35. (New) The system of claim 21 wherein the prosody analyzer further comprises an acoustic feature extraction module that extracts acoustic feature data from said spoken audio signal and wherein the alignment module uses said acoustic feature data to perform the aligning.